Numerals in authorial Turkish-language texts and the stylometric analysis

Andrei Zenkov[1,2,*], *Eugene* Zenkov[2] , *Miroslav* Zenkov[2] , and *Larisa* Sazanova[1]

[1]Ural State University of Economics, 8 Marta St., 62, 620144 Ekaterinburg, Russia
[2]Ural Federal University, Mira St. 19, 620002 Ekaterinburg, Russia

Abstract. Two approaches to the statistical analysis of texts are suggested, both based on the study of numerals occurrence in coherent texts. The first approach is related to the study of the frequency distribution of various leading digits of numerals occurring in the text. These frequencies are unequal: the digit 1 is strongly dominating; usually, the incidence of subsequent digits is monotonically decreasing. The frequencies of occurrence of the digit 1, as well as, to a lesser extent, the digits 2 and 3, are usually a characteristic author's style feature, manifested in all (sufficiently long) texts of any author. This approach is convenient for testing whether a group of texts has common authorship: the latter is dubious if the frequency distributions are sufficiently different. The second approach is the extension of the first one and requires the study of the frequency distribution of numerals themselves (not their leading digits). The approach yields non-trivial information about the author, stylistic and genre peculiarities of the texts and is suited for the advanced discourse analysis. This paper deals with the application of the second approach to the literary texts in Turkish. We have analysed almost the whole corpus of works by are illustrated by examples of computer analysis of the literary texts by O. Pamuk and Y. Kemal – two of Turkey's most prominent novelists. The hierarchical cluster analysis based on the occurrence of numerals in the texts by Pamuk and Kemal shows the author, genre, and chronology differences of numerals usage in the literary texts of these authors.

1 Introduction

The scope of this research relates to stylometry (statistical study of texts in order to find individual features of the author's style – in particular, for attribution of texts). The conventional methods used – taking into account the frequency of occurrence of certain words and collocations in the text, the average length of words and sentences, etc. [1] – often lead to contradictory results, and the very abundance of methods indicates a lack of reliability of each of them individually. In this case, the emergence of new stylometric techniques is not redundant, and they are all complementary rather than mutually exclusive.

We have proposed the idea of studying numerals found in the text as a means of characterizing the author's style [2–6]. The analysis of numerals has many advantages. The

* Corresponding author: zenkow@mail.ru

results of this analysis allow direct linguistic interpretation (unlike, for example, the neural network method [1], which successfully recognizes the texts authorship, but the recognition procedure is a *black box*). The use of numerals in the text is directly related to its authorship, style, and genre (see below).

The approach to stylometry problems we are developing has two varieties. First, we studied the frequency distribution of the leading digits of numerals. The idea may seem strange, but it is in line with research related to Benford's Law [7] – a mysterious, not fully understood manifestation of the Law of large numbers, according to which in large arrays of numerical data describing various objects and phenomena, numbers starting with digit 1 (their share according to Benford's Law is 30.1 per cent) are more common than those starting with digit 2, and the latter, in turn, are more common than numbers starting with digit 3, and so on. According to our research, the leading digits of numerals in coherent texts are distributed even more unevenly than prescribed by Benford's Law: the proportion of numerals starting with 1 can reach 50 per cent. Usually, the frequency distribution of the leading digits of numerals is characteristic of each author and appears in all (large enough) of his works. Sometimes this allows to check the authorship of texts: if the distributions of the leading digits significantly differ for two texts, then the same authorship of the texts is doubtful.

The second variation of our stylometric method consists in analysing the numerals contained in the text (and not their leading digits). The frequency distribution of numerals is also, to a large extent, specific for the author [4–6]. The first of the two approaches can be considered a convolution of the second. Each approach has its own advantages and disadvantages.

Counting the leading digits makes sense only for significant digits 1, 2, and, possibly 3, since the occurrence of subsequent digits is subject to strong fluctuations even in the texts of the same author. Thus, only a small part of the statistical information about the numerals contained in the text is available for analysis. In addition, a problem arises with texts in languages in which the numeral one is formally indistinguishable from the indefinite article (although this is surmountable by switching to an intermediary language without this problem). On the other hand, the information here is presented in a generalized form, which allows to average specific features of individual works of the author.

Analysis of the use of the numerals themselves (not the leading digits) provides richer information about the author's features of the text and, to a large extent, is devoid of indistinguishability of the numeral one and the indefinite article. However, analysing the statistics of numerals is technically more difficult. This article is devoted to the possibilities and comparison of both types of our method.

So far, we have applied this analysis to connected literary texts in Russian, English, Czech, and Lithuanian. In this paper, for the first time we consider the application of our method to Turkish – a language *not* pertaining to the Indo-European family of languages.

2 Objects of research

We studied the literary texts by Orhan Pamuk and Yaşar Kemal – two of Turkey's most prominent novelists, the first of whom is the recipient of the 2006 Nobel Prize in literature, and the second one had been a candidate for that Prize.

We have analysed the following novels by Pamuk:

1) *Cevdet Bey ve Oğulları* (Cevdet Bey and His Sons, 1982),
2) *Sessiz Ev* (Silent House, 1983),
3) *Beyaz Kale* (The White Castle, 1985),
4) *Kara Kitap* (The Black Book, 1990),
5) *Yeni Hayat* (The New Life, 1994),

6) *Benim Adım Kırmızı* (My Name is Red, 1998),
7) *Kar* (Snow, 2002),
8) *Masumiyet Müzesi* (The Museum of Innocence, 2008),
9) *Kafamda Bir Tuhaflık* (A Strangeness in My Mind, 2014),
10) *Kırmızı Saçlı Kadın* (The Red-Haired Woman, 2016).

We have also analysed the following non-fiction texts by Pamuk:

11) *İstanbul: Hatıralar ve Şehir* (Istanbul: Memories and the City; memoirs, 2003),
12) *Babamın Bavulu* (My Father's Suitcase; the Nobel lecture, 2007),
13) *Manzaradan Parçalar: Hayat, Sokaklar, Edebiyat* (Pieces from the View: Life, Streets, Literature; essays, 2010),
14) *Saf ve Düşünceli Romancı* (Naive and Sentimental Novelist; literary criticism, 2011).

Kemal's novels analysed are:

15) *İnce Memed* (Slim Memed, 1955),
16) *Yer Demir Gök Bakır* (Iron Earth, Copper Sky, 1963),
17) *Ölmez Otu* (The Undying Grass, 1968),
18) *Akçasazın Ağaları/Demirciler Çarşısı Cinayeti* (The Agas of Akchasaz Trilogy/Murder in the Ironsmiths Market, 1974),
19) *Akçasazın Ağaları/ Yusufçuk Yusuf* (The Agas of Akchasaz Trilogy/Yusuf, Little Yusuf, 1975),
20) *Yılanı Öldürseler* (To Crush the Serpent, 1976),
21) *Allahın Askerleri* (God's Soldiers, 1978),
22) *Kuşlar da Gitti* (The Birds Have Also Gone, 1978),
23) *Fırat Suyu Kan Akıyor Baksana* (Look, the Euphrates is Flowing with Blood, 1997),
24) *Karıncanın Su İçtiği* (Ant drinking Water, 2002).

3 Method of research

We have prepared a computer program that searches in the Turkish text for numerals expressed both in numbers and verbally. The texts analysed were pre-cleaned of numerals that do not reflect the author's creative intent (numbering of pages, chapters, etc.) or accidentally included in idioms. Since the analysed texts have different sizes, correction coefficients were used to equalize the results on the occurrence of numerals.

Information about numerals found in texts was systematized using the hierarchical cluster analysis [8]. The farthest neighbour clustering was used (which exaggerates differences but provides clearly defined clusters). The smaller the difference in the occurrence of the same numbers in two texts, the greater the similarity (the smaller the "distance" ρ) between these texts, so the Manhattan metric was used

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n} |x_i - y_i|, \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are n-dimensional vectors whose components are the absolute frequency of occurrence of the first n natural numbers found in both analysed texts. In this paper, we assumed $n = 20$.

4 Results and discussion

Based on the frequency distributions of numerals in the works by Pamuk and Kemal, we performed clustering and built a dendrogram (Fig. 1). The numbers to the left of the

dendrogram refer to the texts listed above. The horizontal scale indicates the "distance" between clusters in conventional units.

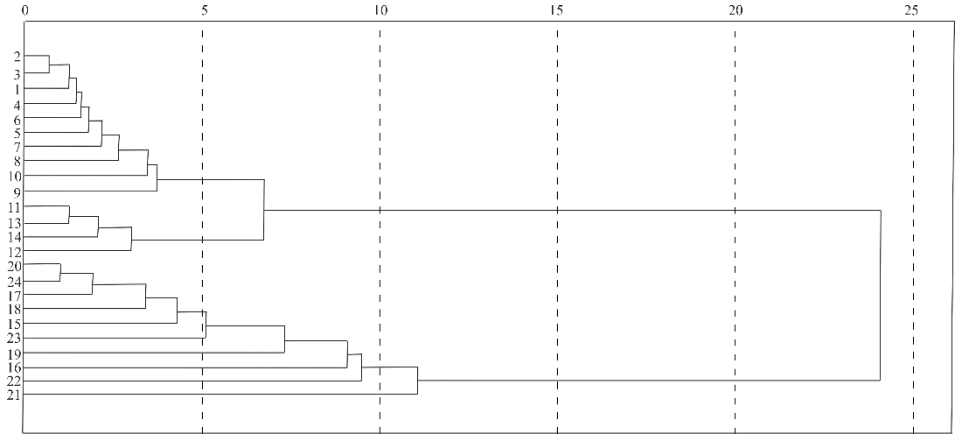

Fig. 1. Results of hierarchical cluster analysis based on the occurrence of numerals in the texts by Pamuk and Kemal. The horizontal scale indicates the "distance" between clusters in conventional units. Texts Nos. 1–24, combined into clusters, are indicated in the article.

Some conclusions directly following from the figure:

1. The genre differences are well marked – cf. the clusters 1–10 and 11–14 containing the fiction and non-fiction works by Pamuk.

2. Pamuk's texts are more uniform in the numerals usage than Kemal's ones.

3. Texts listed above are numbered chronologically – in the sequence of their appearance, and approximately so is the order of their inclusion in the cluster of fiction works by Pamuk. From this, it follows that there is a distinct evolution of numerals usage in his fiction works. In his non-fiction works, we do not see this pattern. This is understandable since the numeral usage in those works is strongly influenced by the work's topic.

4. The cluster of Kemal's (literary) texts is more loose, and his texts lack a strict chronological pattern of numerals usage.

5. The final fusion of all clusters is at a very high level. Pamuk's and Kemal's texts are non-similar in numerals usage

So, the analysis of the use of numerals in texts can be used to test the authorship of texts and distinguish between genres.

5 Conclusions

The analysis shows that taking into account the occurrence of numerals in coherent texts can provide information about the author's, style and genre features of texts. Sometimes, an analysis of the occurrence of numerals allows to reject the hypothesis of the common authorship of texts.

We believe that the methodology we are developing can be a useful addition to the traditional stylometric practices of taking into account the length of sentences and words, the frequency of use of service words and certain significant parts of speech, etc.

This work was supported by a grant from the Russian Foundation for Basic Research, project No. 19-012-00199A, "*A New Method of Text Attribution Based on Statistics of Numerals*". This work was partially supported by a scholarship from the Slovak Academic Information Agency.

References

1. N. Tempestt, S. Kalaivani, F. Aneez, Y. Yiming, X. Yingfei, W. Damon, ACM Comput. Surv., **50**(6), 86 (2017)

2. A. V. Zenkov, Computer Research and Modeling, **9**, 837 (2017)

3. A. V. Zenkov, Journal of Quantitative Linguistics, **25(3)**, 256 (2018)

4. A. V. Zenkov, M. Místecký, Glottometrics, **46**, 12 (2019)

5. A. V. Zenkov, First International Volga Region Conference on Economics, Humanities and Sports (FICEHS 19). Paris, Atlantis Press, Advances in Economics, Business and Management Research, **114**, 448 (2019)

6. A. Zenkov, E. Zenkov, A. Belke, SHS Web of Conferences, **93**, 03026 (2021)

7. F. Benford, Proceedings of American Philosophical Society, **78(4)**, 551 (1938)

8. G. Gan, C. Ma, J. Wu, *Data Clustering: Theory, Algorithms, and Applications* (2007)